# A Multi-modal Perception based Architecture for a Non-intrusive Domestic Assistant Robot*

Christophe Mollaret[1,2,3], Alhayat Ali Mekonnen[2], Julien Pinquier[1], Frédéric Lerasle[2,3], Isabelle Ferrané[1]
[1]IRIT, Univ de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex, France
[2]CNRS, LAAS, 7 avenue du Colonel Roche, F-31400 Toulouse, France
[3]Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France
{cmollare,aamekonn,lerasle}@laas.fr, {pinquier,ferrane}@irit.fr

***Abstract*—We present a multi-modal perception based architecture to realize a non-intrusive domestic assistant robot. The realized robot is non-intrusive in that it only starts interaction with a user when it detects the user's intention to do so automatically. All the robot's actions are based on multi-modal perceptions, which include: user detection based on RGB-D data, user's intention-for-interaction detection with RGB-D and audio data, and communication via speech recognition. The utilization of multi-modal cues in different parts of the robotic activity paves the way to successful robotic runs.**

## I. Introduction

In recent years, autonomous robot usage in hospitals, laboratories, and domestic homes has been increasing. In fact, robots serving various tasks and purposes in the social care and medical/health sectors beyond the traditional scope of surgical and rehabilitation robots are poised to become one of the most important technological innovations of the $21^{st}$ century [2]. In this context, the main issue of Human-Robot Interaction (HRI) research is to endow the robot with suitable behaviors. A reactive behavior, which cycles through monitor and interaction phases has been widely accepted. The interaction generally starts when the user explicitly asks something to the robot or shows an interest to do so. In line with this, we focused on a user intention detection mechanism to transition from monitoring to interaction states. We refer this as *user's intention-for-interaction* detection which is close to the concept of engagement.

In this paper we present an overview of our detection system designed to be non-intrusive. While most research works focus on interaction, the phase before it is also taken into account in our investigation. Following a use-case scenario, experiments involving elderly people have been carried out, giving preliminary evaluations that highlight the soundness of the proposed framework.

## II. Complete System Overview

The overall envisaged framework is illustrated in Fig. 1. This framework is very generic and can be applied to a whole range of scenarios involving HRI. It can be summarized as: detect the person in the environment; go to a garage position and wait until detecting the *user's intention for interaction* based on the fusion of multimodal percepts (e.g., name calling, directing gaze, etc.); then approach this person and begin envisaged close interaction.

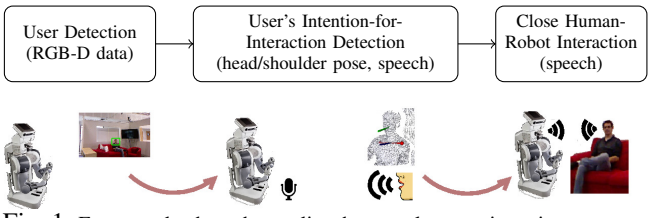


Fig. 1: Framework adopted to realize the complete non-intrusive autonomous assistive robotic system.

*a) User Detection:* Primarily relying on an embedded RGB-D sensor, we use a state-of-the-art upper body detector (head and shoulder), recently proposed by Jafari *et al.* [5], coupled with an optimized HOG based full body detector. The combined detector can detect users that are sitting or standing.

*b) Intention-for-Interaction Detection:* The user's intention is estimated based on three important cues: user's line of sight – inferred from head pose [3]; user's anterior body direction – determined from shoulder orientation computed from openNI skeleton fitting [4]; and speech used to draw attention – captured via Voice Activity Detection (VAD). The head pose detection and shoulder orientation detection modules rely on depth image. Detection outputs are further filtered with a particle swarm optimization inspired tracker developed for this framework. Both, the VAD and tracker output are used as observation inputs and are fused to provide a probabilistic intention estimate using an HMM.

*c) Close Human-Robot Interaction:* During this phase, the human and robot engage in a verbal interaction. The person asks assistance, and the robot answers by providing a useful response or assistance. It is implemented via a static state machine dialogue module that monitors the interaction. Each specific question/request coming from the user can trigger transitions to different states leading to robotic service provision. It uses the CMU PocketSphinx and Google Speech APIs for speech recognition, and the Google TTS API for synthesis. Sentences the robot could say are defined *a priori*.

*d) Task-level Coordination:* Each component of the framework is conceptualized as an individual robotic task that can be coordinated to create and launch a complete working robotic demo. This complex system coordination is done using SMACH [1] – a task-level architecture for rapidly creating complex robot behavior.

*This work was supported by a grant from the French National Research Agency (ANR) under RIDDLE project, grant number ANR-12-CORD-0003.

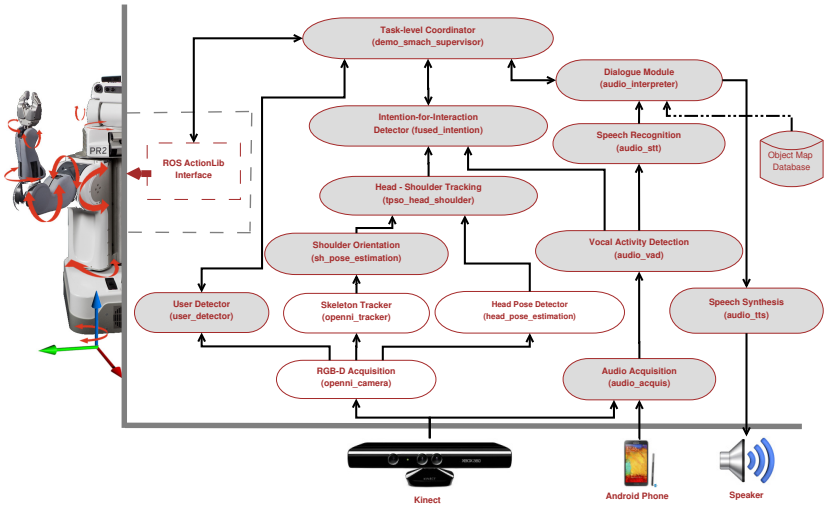


Fig. 2: Illustration of the complete implemented system based on the ROS framework. Each rounded rectangle represents a standalone ROS node and the arrows indicate the message (data) passing pipeline. Shaded nodes correspond to nodes we implemented, the rest denote publicly available implementations.

## III. Implementation and Experiments

(a) H/R situation (b) Kinect view (c) State machine

Fig. 3: Sample snapshots taken during testing with user.

All integration and experiments are carried out on the PR2 robot from Willow Garage Inc. Audio data is captured using Samsung Galaxy Note 2 smartphone (Android 4.2). The smartphone communicates with PR2 via a common wifi network. For experimental purposes, though the framework is quite generic, a specific scenario is realized for assisting elderly people in locating forgotten or misplaced everyday object. Prior to this work, these object positions are recorded by the robot during a "Home-Tour" type scenario in order to build a semantic map of the environment. Fig. 2 depicts the complete software implementation based on the Robotic Operating System (ROS) framework that realizes the different system components.

The complete perception driven architecture integrated on the PR2 robot has been tested with the help of 11 elderly people (ages ranging from 60 to 70 years, with only 7 "robot-familiar" users) – Fig. 3 shows a sample instance. At this stage, the evaluations are targeted towards assessing system soundness with one run per user. An experiment is labeled as successful, if the SMACH based state machine is traversed correctly leading to a "succeeded" output at the end, and it is considered a failure, if by any means, it resulted into an "aborted" or a "preempted" states. If a "succeeded" output has been emitted, the robot successfully answered the user's requests and was sent back to its garage position. In all the cases the robot managed to correctly detect the user, transition to its garage state, detect user's intention, and carry-out the close interaction phase as planned—a 100% mission success rate. In 72.7% of the cases, the robot detected the user's intention-for-interaction on the first user attempt, while in 18.2% of the time it detected it on the second attempt, and the rest on the third attempt, making this a reliable framework. During the interaction, the objects were generally found within 5 exchanges. Hence, the system completely meets expectations and reflects evaluation results obtained during each perceptual component evaluation (though it is not detailed here). No differences were showed between the "robot-familiar" population and the others.

## IV. Conclusions and Future Works

In conclusion, a multi-modal perception based architecture for a non-intrusive domestic assistant robot has been presented. The described system exhibits a non-intrusive characteristic as it engages in close HRI phase only when the user wishes to interact by expressing his/her intent. It relies on a multi-modal user detector, based on RGB-D data, to localize the user in the scene; a multi-modal user's intention-for-interaction detector, based on RGB-D data and VAD; and various ASR APIs for reliable communication. Each perceptual component, though it could not be detailed here, has been evaluated separately leading to a successful non-intrusive robotic system trial runs. These components, through the use-case scenario was carried out on a majority of non "robot-familiar" users, handled a large variety of Human-Robot situations.

In the near future, the presented system will be tested with various users to do a statistical user satisfaction study among others improvements.

## References

[1] J. Bohren. Wiki: smach. http://wiki.ros.org/smach, 2010.
[2] T.-S. Dahl and M.-N. Kamel Boulos. Robots in health and social care: A complementary technology to home care and telehealthcare? *Robotics*, pages 1–21, 2013.
[3] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random forests for real time 3D face analysis. In *IJCV*, 2011.
[4] Tim Field. Openni tracker ROS package (groovy). http://wiki.ros.org/openni_tracker/, 2013.
[5] O. Hosseini Jafari, D. Mitzel, and B. Leibek. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *Int. Conf. on Robotics and Automation (ICRA'14)*, 2014.